CLAIMS

What is claimed is:

1. A data analysis system, comprising:

a first component that facilitates generation of a first data set related to web page information obtained *via* a communication system; and

a second component that coordinates a data set relating to web page information from at least one distributed resource which interacts with the communication system; the second data set is utilized to refine the first data set.

2. The system of claim 1, the first component comprising an internet web crawler.

3. The system of claim 1, the first component comprising an intranet web crawler.

4. The system of claim 1, the second component further utilized to optimize reception of data from the distributed resources.

5. The system of claim 1, the second component provides a scheduling function to control reception of the second data set from the at least one distributed resource.

6. The system of claim 1, the second component utilized to facilitate communication traffic reduction *via* the communication system by employing a proper set of weak indicator functions representative of the first data set.

7. The system of claim 6, the second component further utilized to randomly select and transmit a weak indicator function selected from the proper set of weak indicator functions to at least one of the distributed resources.

8. The system of claim 1, the second component further utilized to compare the first data set and the second data set to detect spoof data retrieved by the first component.

9. The system of claim 1, the second component further utilized to generate status information about data related to the first data set; the status information transmitted to at least one distributed resource.

10. The system of claim 9, the status information comprising, at least in part, a freshness flag to indicate freshness of information related to the first data set.

11. The system of claim 9, the status information comprising, at least in part, a hash of contents of information related to the first data set.

12. The system of claim 9, the status information comprising, at least in part, a copy of information of the first data set.

13. The system of claim 1, the communication system comprising an internet.

14. The system of claim 1, the communication system comprising a world wide web.

15. The system of claim 1, the communication system comprising an intranet.

16. The system of claim 15, the intranet comprising a local area network.

17. The system of claim 15, the intranet comprising a wide area network.

18. The system of claim 1, the distributed resources comprising clients of a server.

19. The system of claim 1, the distributed resources comprising trusted entities interactive with the communication system and the second component.

20. The system of claim 1, the first data set comprising internet web page data.

21. The system of claim 1, the first data set comprising intranet web page data.

22. The system of claim 1, the second data set utilized to add additional data to the first data set; the additional data comprising data unknown to the first component.

23. The system of claim 1, the second data set comprising, at least in part, a hash of contents of at least one web page.

24. The system of claim 1, the second data set comprising, at least in part, a Uniform Resource Locator (URL) of at least one web page.

25. The system of claim 1, the second data set comprising, at least in part, a time stamp relating to an acquisition time for information about at least one web page.

26. The system of claim 1, the second data set comprising, at least in part, a delta indication of changes to contents of at least one web page.

27. The system of claim 26, the delta indication including, at least in part, a hash of previous contents of a web page and a hash of recent contents of the web page.

28. The system of claim 1, the second data set comprising, at least in part, a status indication of changes to contents of at least one web page.

29. The system of claim 28, the status indication including, at least in part, a percentage relating to an amount of change of contents of a web page.

30. The system of claim 28, the status indication including, at least in part, a significance indicator to signify importance of changes in contents of a web page.

31. The system of claim 1, the second data set comprising internet web page data.

32. The system of claim 1, the second data set comprising intranet web page data.

33. The system of claim 1, the second data set comprising data compiled utilizing at least one weak indicator function randomly selected from a set of weak indicator functions; the set of weak indicator functions representative of the first data set.

34. The system of claim 1, further comprising a search component to accept at least one search query and generate at least one search reply having at least a portion of the first data set represented by information embedded in the search reply.

35. The system of claim 1, further comprising a web page server component to construct web pages having at least a portion of the first data set

represented by information embedded in at least one link found on at least one constructed web page.

36. The system of claim 1, further comprising a storage component to store the first data set.

37. A method for facilitating data analysis, comprising:

generating a first data set relating to a second data set obtained from web pages interactive with a communication system;

receiving a third data set from at least one distributed resource that is interactive with the communication system; the third data set comprising web page related information generated by the distributed resource; and

refining the second data set to reflect information obtained from the third data set.

38. The method of claim 37, the first data set comprising a representation of the second data set.

39. The method of claim 38, the representation of the second data set comprising, at least in part, a hash of contents of at least one web page contained in the second data set.

40. The method of claim 38, the representation of the second data set comprising, at least in part, a status indication of at least one web page contained in the second data set.

41. The method of claim 40, the status indication comprising a freshness flag to indicate if the web page information is current.

42. The method of claim 37, the first data set comprising a copy of the second data set.

43. The method of claim 37, the second data set comprising web page information compiled by a web crawler.

44. The method of claim 37, the third data set comprising web page information based upon client accessed web page information on the communication system.

45. The method of claim 37, the distributed resource comprising a client of a distributed crawler system.

46. The method of claim 37, the communication system comprising an internet.

47. The method of claim 37, the communication system comprising an intranet.

48. The method of claim 37, refining the second data set comprising:

adding unknown information to the second data set when new information is received from the distributed source *via* the third data set;

updating existing information in the second data set when changes have occurred as indicated by the third data set; and

resetting any indicators utilized to pass status information to the distributed resources after information from the third data set has been analyzed.

49. The method of claim 37, further including:

transmitting the first data set to at least one distributed resource that is interactive with the communication system making the first data set available to be utilized by the distributed resource to generate the third data set.

50. The method of claim 38, further including:

generating a set of weak indicator functions to represent the second data set; and

selecting random weak indicator functions from the set of weak indicator functions to transmit to the distributed resources as the first data set.

51. The method of claim 50, the set of weak indicator functions comprising a proper set of weak indicator functions such that a non-zero probability exists that a randomly selected weak indicator function can identify a new web page.

52. The method of claim 50, generating a set of weak indicator functions comprising:

providing a dictionary representative of the second data set;

partitioning randomly the dictionary into non-overlapping subdictionaries; and

creating a function where $I(x) = 1$ if and only if at least one subdictionary's weak indicator function is equal to one.

53. The method of claim 37, further including:

comparing the third data set to the second data set to reveal spoof data included in the second data set.

54. The method of claim 37, further including:

optimizing reception of at least one third data set through scheduling of the distributed resources.

55. The method of claim 37, further including:

receiving a web page search query from at least one distributed resource;

generating a web search results page in response to the web page search query from the distributed resource;

embedding portions of the first data set in links found on the web search results page; and

transmitting the web search results page as a representation of at least a portion of the second data set to the distributed resource.

56. The method of claim 37, further including:

constructing a web page utilizing at least a portion of the first data set to embed information about links found in the web page; and

transmitting the web page to disseminate the first data set to at least one distributed resource.

57. A data analysis system, comprising:

means for generating at least one first data set from a communication system;

means for receiving and coordinating at least one second data set from at least one distributed resource which interacts with the communication system; and

means for refining the first data set utilizing at least one second data set.

58. The system of claim 57, the means for generating at least one first data set including a web crawler.

59. The system of claim 58, the first data set comprising data relating to web pages obtained by the web crawler.

60. The system of claim 57, the second data set comprising web page comparison data compiled by at least one distributed resource and based, at least in part, upon representative data of the first data set.

61. A data analysis system, comprising:

a first component that generates web page information from at least one visited web site for utilization in a distributed web crawling system; the web page information transmitted by the first component to a second component *via* a communication system.

62. The system of claim 61, the first component providing at least one time stamp relevant to a time of acquisition of data utilized in the generation of the web page information.

63. The system of claim 61, the first component receiving a set of embedded web crawler data from at least one search result page to utilize in the generation of the web page information.

64. The system of claim 61, the first component receiving a set of embedded web crawler data from at least one web page to utilize in the generation of the web page information.

65. The system of claim 61, the first component further operational to obtain web page data indirectly *via* at least one other client of the distributed crawler system to provide a gateway to a second component to substantially reduce traffic flow to the second component.

66. The system of claim 61, the first component receiving web page related data from at least one client and at least one server of the distributed web crawling system.

67. The system of claim 61, the generated web page information comprising, at least in part, a status indication of changes to contents of at least one web page.

68. The system of claim 67, the status indication including, at least in part, a percentage relating to an amount of change of contents of a web page.

69. The system of claim 67, the status indication including, at least in part, a significance indicator to signify importance of changes in contents of a web page.

70. The system of claim 61, at least a portion of the generated web page information made available for peer-to-peer client transmission *via* the communication system.

71. The system of claim 61, the generated web page information compiled utilizing a randomly selected weak indicator function from a proper set of weak indicator functions that represent web page data compiled by a web crawler.

72. The system of claim 61, the communication system comprising an internet.

73. The system of claim 61, the communication system comprising an intranet.

74. The system of claim 61, further comprising a storage component to store the web page information.

75. The system of claim 61, further comprising a notification component that determines when and if the generated web page information is to be communicated *via* the communication system.

76. The system of claim 75, the notification component receiving scheduling information from a second component; the scheduling information relating to obtaining and transmitting the generated web page information.

77. The system of claim 61, the first component receiving a set of data from a second component to utilize in the generation of the web page information.

78. The system of claim 77, the first component utilizing web search servers outside of the distributed web crawling system to retrieve data unknown to the second component.

79. The system of claim 77, the first component generates comparison data based on the web page information and the received set of data; the first

component making the comparison data discretionarily available to the second component *via* the communication system.

80. The system of claim 79, the comparison data including, at least in part, at least one Uniform Resource Locator (URL) of at least one web page.

81. The system of claim 79, the comparison data including, at least in part, a hash of contents of at least one web page representative of a recent web site visit.

82. The system of claim 79, the comparison data including, at least in part, a delta indication of contents of at least one web page.

83. The system of claim 82, the delta indication including, at least in part, a hash of previous contents of a web page and a hash of recent contents of the web page.

84. The system of claim 77, the second component comprising a server of the distributed crawling system.

85. The system of claim 77, the second component comprising a client of the distributed crawling system.

86. The system of claim 77, the generated web page information comprising data unknown to the second component.

87. The system of claim 77, at least a portion of the received set of data made available for peer-to-peer client transmission *via* the communication system.

88. The system of claim 77, the received set of data comprising a dictionary for data compiled by a web crawler.

89. The system of claim 77, the received set of data comprising a representation of data compiled by a web crawler; the representation of data generated by utilizing a weak indicator function.

90. The system of claim 77, the received set of data comprising a copy of data compiled by a web crawler.

91. The system of claim 77, further comprising a storage component to store the set of data received from the second component.

92. A method for facilitating data analysis, comprising:

compiling a first data set derived from accessing web pages *via* a communication system; and

transmitting, selectively, the first data set to an entity of a distributed crawling system that is interactive with the communication system.

93. The method of claim 92, the entity comprising a server of the distributed crawling system.

94. The method of claim 92, the entity comprising at least one client of the distributed crawling system.

95. The method of claim 92, the first data set comprising, at least in part, a uniform resource locator (URL) for at least one web page.

96. The method of claim 92, the first data set comprising, at least in part, a hash of contents of at least one web page.

97. The method of claim 92, selectively transmitting based upon time of day.

98. The method of claim 92, selectively transmitting based upon priority of at least one web page.

99. The method of claim 92, selectively transmitting based upon percentage of content change of at least one web page.

100. The method of claim 92, selectively transmitting based upon identifying at least one new web page.

101. The method of claim 92, further comprising:

receiving a representation of a second data set compiled by a web crawler; the second data set relating to at least one web page from the communication system.

102. The method of claim 101, receiving the representation of the second data set is accomplished *via* reception of a web page with embedded information derived from the second data set and generated by a web page hosting server with access to the second data set.

103. The method of claim 101, receiving the representation of the second data set is accomplished *via* reception of a search results page with embedded information derived from the second data set and generated in response to a query transmitted to a search server having access to the second data set.

104. The method of claim 101, further comprising:

utilizing the second data set to control which web pages to visit to compile the first data set.

105. The method of claim 101, further comprising:

determining when to transmit the first data set *via* the communication system based upon the second data set.

106. The method of claim 105, the second data set containing a freshness indicator to indicate when its data is stale and requires updating *via* the first data set.

107. The method of claim 105, the second data set containing a schedule for when the first data set is to be transmitted.

108. The method of claim 101, further comprising:

comparing at least a portion of the second data set with at least a portion of information obtained *via* accessing web pages to create comparison data; and

generating a representation of the comparison data to derive the first data set.

109. The method of claim 108, the first data set comprising data unknown to the second data set.

110. The method of claim 109, the unknown data comprising only unknown data derived from at least one search results page from a search server outside of the distributed crawling system.

111. The method of claim 108, the first data set comprising content changes to web pages represented by the second data set.

112. The method of claim 108, the first data set comprising status information relating to web pages represented by the second data set.

113. A data packet transmitted between two or more computer components that facilitate information gathering, the data packet comprising, at least in part, information relating to web crawling that utilizes, at least in part, a distributed system for gathering information about web pages.

114. A computer readable medium having stored thereon computer executable components of the system of claim 1.

115. A device employing the method of claim 37 comprising at least one selected from the group consisting of a computer, a server, and a handheld electronic device.

116. A device employing the system of claim 1 comprising at least one selected from the group consisting of a computer, a server, and a handheld electronic device.